

FLASH MEMORY DEVICE HAVING A PAGE MODE OF OPERATION

This is a continuation of application Ser. No. 08/699,490, filed Aug. 19, 1996, now abandoned.

FIELD OF THE INVENTION

The present invention relates generally to memory devices and more particularly to a nonvolatile memory device having a page mode of operation.

BACKGROUND OF THE INVENTION

Nonvolatile memory devices such as Electrically Programmable Read Only Memories ("EPROMs"), Electrically Erasable Programmable Read Only Memories ("EEPROMs"), and flash EEPROMs include an array of nonvolatile memory cells and supporting periphery circuitry for accessing the array. A nonvolatile memory cell typically behaves like a field effect transistor and includes a select (or control) gate that controls the reading and writing of data to the memory cell and a floating gate that traps charge to alter the datum or data stored by the memory cell.

As charge is added to the floating gate of a memory cell, the threshold voltage V_t of the memory cell increases, and the memory cell drain current I_D ("cell current") decreases. The memory cell threshold voltage V_t is related to the memory cell drain current I_D by the expression:

$$I_D \propto G_m (V_G - V_t) \text{ for } V_D > V_G - V_t$$

wherein G_m is the transconductance of the memory cell; V_G is the memory cell gate voltage; V_D is the memory cell drain voltage; and V_t is the memory cell threshold voltage.

Given this relationship, there are a number of prior art methods for sensing the amount of charge stored on the floating gate of the memory cell, including the following:

- 1) sensing the cell current of a memory cell when a constant voltage is applied to the select gate of the memory cell;
- 2) sensing the amount of voltage required at the select gate to give rise to an expected cell current for the memory cell;
- 3) sensing a voltage drop across a load that is coupled to the drain of the memory cell when a constant voltage is applied to the select gate of the memory cell, wherein the cell current determines the amount of the voltage drop across the load; and
- 4) sensing the amount of voltage required at the select gate to give rise to an expected voltage drop across a load that is coupled to the drain of the memory cell.

Once sensed, the amount of charge determined to be stored on the floating gate is decoded to correspond to one of n possible states, n being two or more, and the binary representation ($\log_2 n$) of the determined state is output. One disadvantage of the above-described methods, all of which require an active cell current, is that a relatively large amount of current is required for sensing each cell, which reduces the maximum number of cells that may be sensed in parallel.

FIG. 1 shows a prior sensing system 5 that is used to sense the state of, and therefore the data stored by, nonvolatile memory cell 10. Nonvolatile memory cell 10 includes a select gate SG, a floating gate FG, a source S, and a drain D. Memory cell 10 operates as a field effect transistor (FET) having a variable threshold voltage V_t that is changed by adding and removing charge from the floating gate FG.

Because memory cell 10 operates as a field effect transistor, the electrodes shown in FIG. 1 as the drain D and the source S may interchangeably be used as either source or drain depending upon the particular configuration of memory cell 10 and the operating characteristics applied thereto.

As shown, the prior art sensing system 5 detects the amount of cell drain current I_D that results from applying a read voltage V_G to the select gate SG of memory cell 10. Depending upon the amount of charge stored on the floating gate FG of memory cell 10, the cell current I_D may vary anywhere from zero to approximately 100 microamperes when the read voltage V_G is applied to select gate SG.

The select gate SG of memory cell 10 is coupled to a wordline (not shown) to receive the read voltage V_G , the drain of memory cell 10 is coupled to a bitline (not shown) to which sensing system 5 is coupled to detect the strength of the cell current I_D , and the source of memory cell 10 is coupled to a ground potential V_{SS} to give rise to the cell current I_D that flows from the drain to the source as shown. Thus, memory cell 10 operates as a pull-down device.

A corresponding pull-up device is found in column load circuit 19. Column load circuit 19 is shown as including a transistor 20 that is biased to operate as a pull up device by a gate voltage V_{bias} . A drain bias circuit 12 is coupled between the drain D of memory cell 10 and column load circuit 19 to ensure that the drain D of memory cell 10 does not drop below a predetermined voltage (e.g. approximately one volt). Drain bias circuit 12 is shown as including a cascode transistor 15 and feedback circuitry 17. Feedback circuitry 17 provides a necessary voltage to the gate of transistor 15 such that the drain D of memory cell 10 does not drop below the predetermined voltage.

Once the read voltage V_G is applied to the select gate SG of memory cell 10, the amount of charge trapped on floating gate FG determines the strength of the cell current I_D and the strength of the pull-down provided by memory cell 10. Typically, if the memory cell 10 is erased, memory cell 10 acts as a strong pull-down device to overcome the pull-up provided by column load circuit 19 such that the negative input of a differential sense amplifier 25 is discharged towards ground. Sense amplifier 25 compares the voltage at its negative input to the voltage at its positive input, which is supplied by a reference circuit 30. According to common prior techniques, reference circuit 30 includes a reference cell (not shown) that has its floating gate charged to a predetermined level coupled to a drain bias circuit and column load circuit identical to those shown of sensing system 5.

Using the sensing scheme embodied by sensing system 5 requires a relatively large amount of current to read memory cell 10. For example, each of the memory cell 10, the drain bias circuit 12, and the sense amplifier 25 require current for operation and therefore result in power consumption. The amount of power consumption required by these components of sensing system 5 results in the ability of sensing system 5 to sense relatively few memory cells (e.g. 16 or 32) in parallel.

SUMMARY AND OBJECTS OF THE INVENTION

Therefore, it is one object of the present invention to provide a method for sensing more nonvolatile memory cells in parallel than previously allowed by typical prior art methods.

It is a further object of the present invention to use the method to provide a page mode of operation for a nonvolatile memory device.

It is a further object of the present invention to provide a method for sensing the cell threshold of memory cell rather than the cell current of the memory cell.

It is a further object of the present invention to provide a method for sensing a memory cell without using a drain bias circuit or a differential sense amplifier.

A method for determining data stored by a memory cell is described. The memory cell has a select gate coupled to a wordline, a first electrode coupled to a bitline, and a second electrode coupled to a conductor. The method comprises: floating the bitline; applying a first voltage to the wordline; applying a second voltage to the conductor such that the bitline is set to a third voltage that is equal to the first voltage minus a threshold voltage of the memory cell; and sensing the third voltage to determine the data stored by the memory cell. For one embodiment, the memory cell is a nonvolatile memory cell.

Other objects, features, and advantages of the present invention will be apparent from the accompanying drawings and from the detailed description which follows below.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example and not limitation in the figures of the accompanying drawings, in which like references indicate similar elements, and in which:

FIG. 1 is prior art circuit diagram of a sensing system including a memory cell, a drain bias circuit, a column load circuit, and a differential sense amplifier.

FIG. 2 is a nonvolatile memory cell configured according to one steps of a method for sensing data stored in the nonvolatile memory cell;

FIG. 3 is the nonvolatile memory cell of FIG. 2 configured according to another step of the method for sensing data stored in the nonvolatile memory cell;

FIG. 4 is the nonvolatile memory cell of FIG. 2 configured according to another step of the method for sensing data stored in the nonvolatile memory cell;

FIG. 5 is the nonvolatile memory cell of FIG. 2 configured according to another step of the method for sensing data stored in the nonvolatile memory cell;

FIG. 6 is a programming window for a nonvolatile memory cell;

FIG. 7 is a memory device including a memory cell array and periphery circuitry;

FIG. 8 is one embodiment of a memory cell array of FIG. 7 including wordlines, bitlines, source straps, source diffusion, and memory cells;

FIG. 9 is a method for sensing data stored in memory cells coupled to a wordline; and

FIG. 10 is one embodiment of periphery circuitry of FIG. 7 including a voltage regulation circuit, voltage switches, row and column decoders, sensing circuitry, and an optional control engine.

DETAILED DESCRIPTION

A method and associated apparatus are described herein that provide for sensing the data stored by an entire wordline of a memory device in parallel without the excessive consumption of current. The advantages of the present described method and apparatus are provided, in part, by performing read operations of the memory cells without requiring an active cell current I_D (i.e. approximately zero DC cell current). Because no cell current is required to sense the data

stored by the nonvolatile memory, the amount of power required to read each cell is reduced sufficiently to allow for an entire selected wordline of memory cells to be read simultaneously. Wherein the amount of time required to form a single read operation using the described methods may, under some circumstances, be longer than the amount of time required to read a memory cell using some prior art schemes, the fact that many more memory cells can be sensed in parallel allows the sensing time to be amortized over many more cells so that the average amount of time to read each cell is significantly reduced and performance is increased.

The sensing system that operates according to present embodiments, as described with respect to FIGS. 2-10, does not require the drain bias circuit, a differential sense amplifier, or an active cell current in order to accurately sense the amount of charge stored on the floating gate of a memory cell. Instead, the sensing system and method described herein uses a source bias and an analog-to-digital converter (ADC) to detect the state of a cell in a manner that will be described below. FIGS. 2-5 diagrammatically illustrate the manner in which the state of a memory cell may be sensed according to the present embodiments. FIG. 2 shows a memory cell 35 having its select gate SG coupled to a wordline 40, its "drain" D coupled to a bitline 45, and its source S floating. The process of sensing the state of memory cell 35 is begun in FIG. 2 by setting the voltage of bitline 45, and therefore the "drain" D of memory cell 35, to system ground V_{SS} . Alternatively, the drain D of memory cell 35 may not first be set to V_{SS} . As shown in FIG. 3, the ground potential V_{SS} is removed from bitline 45 such that both the "drain" D and "source" S of memory cell 35 are floating. As shown in FIG. 4, a read voltage V_G is then applied to wordline 40 while a bitline 45 and the source of memory cell 35 remain floating.

In FIG. 5, a source voltage V_S is applied to the "source" S of memory cell 35, which results in the "drain" D of memory cell 35 being pulled up to a voltage proportional to $(V_G - V_t)$. The value of the voltage of the bitline 45 may be sensed, for example, using an analog to digital converter.

Applying the source voltage V_S to memory cell 35 actually results in the functions of the "drain" and "source" of memory cell 35 reversing such that what was formerly the "drain" of memory cell 35 now operates as the "source" of memory cell 35, and vice versa.

The method and system shown in FIGS. 2-5 enables the actual cell threshold voltage V_t of memory cell 35 to be sensed. Many prior schemes sense the cell current I_D and not the cell threshold voltage V_t of a memory cell. The cell current I_D is a function of many parameters that can lead to sources of variation in the cell current I_D from memory cell to memory cell.

When memory cell 35 is a flash memory cell, the cell threshold voltage V_t determines the cell current I_D . Therefore, in order to reduce the variability of the sensed data stored in memory cells, it is preferable to sense the cell threshold voltage V_t rather than the cell current I_D . Additionally, many of a flash memory cell's parameters rely on the cell threshold voltage V_t , not the cell current I_D . For instance, the programming of a flash memory cell is directly proportional to the cell threshold voltage V_t , not the cell current I_D . In addition, disturb mechanisms (charge loss and charge gain) are related to the cell threshold voltage V_t . Therefore, it is preferable to sense the cell threshold voltage V_t rather than the cell current I_D .

The scheme illustrated in FIGS. 2-5 does not require an active cell current I_D , the use of a drain bias circuit, or a

differential sense amplifier. Therefore, substantially less power is required to sense data stored by memory cell 35 as opposed to prior memory cells (e.g. memory cell 10 of FIG. 1). The reduced power required to sense data stored in memory cell 35 enables more memory cells to be sensed or read in parallel than previous schemes would allow. For one embodiment, an entire wordline of memory cells (e.g. 1024) may be sensed in parallel.

According to the present embodiments, both the read voltage V_G applied to the wordline 40 and the source voltage V_S are selected in view of the operating characteristics of the memory cell 35. Specifically, for the example wherein memory cell 35 is a flash memory cell, the range of possible cell threshold voltages V_t provides a programming window that may be subdivided into a number of distinct states. The number of states into which the programming window is divided determines the number of bits stored by the memory cell 35. For example, if a programming window is subdivided into only two states, memory cell 35 is capable of storing only one bit of data. Alternatively, if a programming window is subdivided into eight distinct states, memory cell 35 is capable of storing 3 bits of data.

FIG. 6 shows a programming window 50 as a function of the threshold voltage V_t . As shown, a minimum threshold voltage V_{tmin} defines a lower boundary of the programming window 50, and a maximum threshold voltage V_{tmax} defines an upper bound of the programming window 50. The lower bound of programming window 50 is determined primarily by over-erased conditions of the flash memory cell, and the upper bound of programming window 50 is determined primarily by maximum programming voltages and cell disturbs.

According to the present embodiments, the read voltage V_G is selected to be greater than the maximum threshold voltage V_{tmax} . The source voltage V_S is selected to be equal to $(V_{tmax} - V_{tmin})$, which is the maximum possible swing of the bitline. When these constant voltages are applied to wordline 40 and the source S of a memory cell 35, bitline 45 coupled to the drain D of memory cell 35 will eventually achieve a DC value proportional to $(V_G - V_t)$. For one embodiment, the read voltage V_G is approximately 5.5 volts and the source voltage V_S is approximately 3.0 volts.

FIG. 7 shows one embodiment of a memory device 70 that includes a memory cell array 75 and periphery circuitry 80. Memory cell array 75 comprises a multiplicity of memory cells arranged in rows (wordlines) and columns (bitlines) such as those shown in FIG. 8. Periphery circuitry 80 includes circuitry for reading and writing of data to the memory cell array 75 and an interface to the data bus 85. Periphery circuitry 80 is shown as including power supply inputs V_{CC} and V_{SS} , programming voltage V_{PP} , control inputs CTL (including chip enable, output enable, etc.), address inputs ADDR, and an I/O path to data bus 85.

FIG. 8 shows one embodiment of memory cell array 75. Depending upon the storage capacity of memory device 70, memory cell array 75 may include several hundreds of bitlines and several hundreds of wordlines. FIG. 8 shows a sub-section of memory cell array 75 that includes a plurality of memory cells 101-103, all of which are coupled to wordline 100, and another plurality of memory cells 111-113, all of which are coupled to wordline 110. Bitline 120 is shown as being coupled to the drains of memory cells 101 and 111, bitline 121 is shown as being coupled to the drains of memory cells 102 and 112, bitline 122 is shown as being coupled to the drains of memory cells 103 and 113. Each of the memory cells 101-103 and 111-113 are shown

as having their sources coupled to a source diffusion 130 that is coupled to a source strap 131. Additional bitlines, wordlines, source diffusions, source straps, and memory cells are implied by FIG. 8.

Memory cell array 75 of FIG. 8 may be manipulated in the manner shown in FIGS. 2-5 by appropriately applying voltages to bitlines 120-122, source straps 131-132, and wordlines 100 and 110. As may be seen, a voltage applied to a particular wordline will be applied to the select gates of all of the memory cells coupled to that wordline. For example, applying a voltage to wordline 100 will result in each of the memory cells 101-103 having that voltage applied to their select gates. As will be discussed with respect to FIG. 10, the manipulation of wordlines 100 and 110, bitlines 120-122, and source straps 131-132 may be done using conventional circuitry.

FIG. 9 summarizes a method for reading all of the memory cells coupled to a single wordline (e.g. wordline 100 of FIG. 8). The process begins at process block 150 typically in response to receiving a read request as decoded from the control signals of memory device 70. At process block 155, all bitlines of the memory cell array 75 are grounded, and the bitlines are floated at process block 160. The actions of process blocks 155 and 160 may be performed concurrently with the decoding of an address by the row and column decoders of memory device 70 (shown in FIG. 10). Given that all of the memory cells coupled to a wordline may be simultaneously sensed, the use of a column decoder may not be required.

At process block 165, the read voltage V_G is applied to the selected wordline, and deselected wordlines are grounded to prevent the switching on of the memory cells attached to the deselected wordlines. Again, the read voltage V_G that is applied to the selected wordline is selected to be greater than the maximum threshold voltage V_{tmax} obtainable by a memory cell of memory cell array 75.

At process block 170, a source voltage V_S is applied to all the sources of all the memory cells in memory array 75 by coupling the source voltage V_S to the source straps of the memory cell array. For other layouts of memory cell array 75, the source voltage V_S may not necessarily be applied to all the sources of the memory cells; however, the present scheme will work so long as all of the memory cells of the memory cell array 75 that are to be read have their sources set to the appropriate source voltage V_S .

A predetermined period of time is allowed to elapse so that the bitlines coupled to the selected memory cells are allowed to achieve a steady state voltage proportional to $(V_G - V_t)$. Each of bitlines 120-122 may have a different voltage as determined by the threshold voltages V_t of each of memory cells 101-103. At process block 175 bitline voltages for each of bitlines 120-122 of the memory cell array 75 are sensed. The bitlines may be sensed in parallel or subsets of the bitlines may be sensed sequentially. For example, given 1024 bitlines, the periphery circuitry 80 of the memory device 70 may be configured to sense the voltages of all the bitlines in parallel or to sense a subset of the bitlines at a time (e.g. sixteen). So long as the read voltage V_G and the source voltage V_S are applied to the memory cells of the selected wordline, the bitline voltages on bitlines 120-122 will maintain a DC value proportional to $(V_G - V_t)$.

Providing sufficient time to allow bitlines 120-122 to achieve a DC voltage may require a longer period of time than normally required by prior art sensing schemes; however, the fact that all of the memory cells of an entire